DOCUMENT RESUME

ED 362 005 FL 021 383

AUTHOR Lee, Tony
TITLE Taking a Multi-Faceted View of the Uni-Dimensional Measurement from Rasch Analysis in Language Tests.
PUB DATE 6 Aug 93
NOTE 10p.; Paper presented at the Annual Meeting of the Language Testing Research Colloquium (15th, Cambridge, England, United Kingdom, August 3, 1993).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *English (Second Language); *English for Academic Purposes; Foreign Countries; Higher Education; *Language Proficiency; *Language Tests; Undergraduate Students
IDENTIFIERS FACETS Model; *Rasch Model

ABSTRACT

The study reported in this paper originated from a need to design a short English for Academic Purposes (EAP) proficiency test for incoming undergraduate students. The approach to solving the problem was to focus on establishing consistent person comparison between the students at Hong Kong Baptist College who did not meet the minimum entry requirement in the English language and a reference group who met the required grade level. The comparison was made on the basis of a short English-as-a-Second-Language (ESL) proficiency test taken by both students groups. The establishment of equivalence between the reference and student groups was achieved through the employment of FACETS (Linacre and Wright, 1990). The logit level of the reference group (-0.93) can be taken as the equivalence of the minimal required ESL level for entry into university. Findings suggest the possible extensions of the Rasch model in terms of both item calibration and person measurement through the employment of FACETS. The EAP Test is appended. (Contains 19 references.) (JP)

Taking a multi-faceted view of the uni-dimensional measurement from Rasch analysis in language tests

Tony Lee

I. Introduction

The advent of Item Response Model (IRM) to the field of language testing (e.g. Henning 1984, Henning et al. 1985, Griffin et al. 1985, Woods & Baker 1985, Pollitt & Hutchinson 1987 and Choi & Bachman 1992) has been a most important development in the recent history of the discipline. IRM has given language testing a rigorous basis for measurement. The catch, though, is that IRM is primarily a measurement model with little or no immediate implications for language testing research. Specifically, the uni-dimensionality assumption in IRM has been an initial stumbling block for many language testers. It is argued that, if language is inherently complex, it would be strait-jacketing language testing research by forcing the uni-dimesional condition onto all language data. (See Bachman 1990 and Henning 1992 for an interesting discussion.)

Conceptual analysis (eg. Reckase 1979, Henning et al. 1985, 1992, Choi & Bachman 1992) has helped to define the scope of the uni-dimensionalty assumption and to resolve the apparent dilemma. Research designs encompassing an IRM component have also been developed; and this has helped the applied linguistic research dimension of IRM.

II. Rasch model as a research tool

Wright & Masters (1982) maintain that the uni-dimensionality assumption is a "universal characteristic of all measurement". This, however, should not in theory preclude analyses over-and-above an IRM analysis. Jensen (1978), for example, warns of "... a flagrant conceptual and scientific blunder ... to orthogonal rotation of principal components or factors without first extracting the general factor (i.e. the first principal component or first principal factor)". Indeed, IRM can easily be conceptualized as a rigorous way to extract the general factor. The standardized residuals from an IRM analysis would provide data for further analyses as envisaged by Jensen. Pollitt (personal communication) suggests using residual analysis to tap specific dimensions within behavioural data after the latent trait has been extracted. Lee (1992) analyzes the residuals to establish the construct validity of an ESL reading test.

From the development within IRM itself, multi-faceted Rasch analysis (Linacre 1989a) is the expansion of the one-parameter Rasch model to encompass analysis of facets in the data. This has enabled IRM to be employed in diverse research design and analysis configurations and data collection schedules.

III. Many-Faceted Rasch Analysis

Linacre (1989 a&b) argues and demonstrates the possbility of extending the initial one-parameter (or two-facet) Rasch model to n-facet models. This is an interesting development. Constituents within a complex human behavioural context can now be accommodated within the same IRM model for analysis. Typically facets can include judges of human performance (eg. in a writing test), or sub-groupings of subjects/candidates, or sub-test item groups. With the flexibility introduced, research designs can now be developed which would do greater justice to features in human behavioural (eg. varying severity of judges, cultural and/or economic background of subjects). In addition, FACETS (Lincare & Wright 1990), which is the software implementation of multi-facetd Rasch analysis can generate interaction analyses of the facets. Lee et al. (forthcoming) uses FACETS to calibrate and to establish the scale structure of the Australian Second Language Proficiency Ratings (ASLPR) (Ingram & Wylie 1979). Lee (in preparation) examines, via FACETS, ESL program entry level and test time interaction, and rater and ethnic background interaction in the

National Languages and Literacy Institute of Australia (NLLIA) ESL Bandscales (MacKay et al. 1992).

IV. The Study

A. The Background

The study reported in this paper originated from a need to design a short EAP proficiency test for incoming mature undergraduate students. A policy decision of the Hong Kong Baptist College in 1992 to admit former non-degree graduates into undergraduate degree programs resulted in a situation where the minimum entry requirement in the English language (Grade D in the Use of English examinations of the Hong Kong Examinations Authority) was not met by some of the admittants. It was difficult for these students to re-take the Use of English examinations and uneconomical to administer a facsimile version. The Language Centre was given the charge to find a means to establish the equivalence of the required minimum ESL level for entry into degree programs. The approach to solving the problem was to focus on establishing consistent person comparison between the students in question and a reference group with the required Grade D level in the use of English Examinations. The comparison was made on the basis of a short ESL proficiency test taken by both groups of candidates.

B. The EAP Test

Practical and monitary constraints necessitated the choice of the modified cloze format based on a single reading passage. Two sets of items were prepared. The first consisted of 52 proofreading items relating to grammatical features in the first part of the passage. The second set consisted of 44 gap-filling items relating to cohesion features.

The test was first piloted on a group of undergraduate students covering the whole range of the Use of English grade levels. The test was then given to 221 mature students. A reference group of students (n = 38) with the required minimum Use of English grade was also given the test.

C. The research design

The overall design of the study was to obtain a comparison between the 'student' group and the 'reference' group based on the EAP test. As the reference group was a sample of those who had achieved the required minimum English language standard for entry into universities those in the 'student' group who would match the level of the 'reference' group in the EAP test had to be considered as having an equivalent level of English language ability.

V. The Analysis

To achieve the objectives described above it was necessary to have an ability scale that was robust and consistent and to make the required comparison of the two groups of candidates beyond the particular EAP test given. This was a typical sample free test calibration and test free person measurement in IRM. In addition, the analysis had to calibrate the two sub-groups of candidates (student and reference). Multi-faceted Rasch analysis was thus necessary. It was also thought relevant to calibrate the two parts of the EAP test to see if mastery of grammar and cohesion features were distinguishable in the EAP test. FACETS (Linacre & Wright 1990) was employed. Four facets were included in the analysis: the candidates, the two candidate sub-group: 'Student' and 'Reference, the two sub-tests and the test items.

VI. Results

A. Uni-dimensionality

An informal test of uni-dimensionality was performed via maximum likelihood factor analysis. A first factor containing 21% of the overall variance was

obtained. This was considered sufficient to make a uni-dimensionality claim for Rasch analysis (Rechase 1979, Henning et al. 1985).

5. Item Calibration

Table 2 contains detailed item calibaration of the EAP test. The leftmost column contains descriptive statistics: 'Score' is the raw score of the item across all candidates; 'Count' is the number of score points and 'Average' the item facility value. The second column contains the item calaibratiion statistics: the logit and its associated standard error. The third column contains the fit statistics. FACETS includes two types of fit statistics: the Infit and the Outfit. The former is an information-weighted mean-square fit statistic and the latter the conventional mean-square. The expected value is 1 in both. Values greater than 1 would indicate noise in the Infit statistic and an oulier in the Outfit statistic.

Score	Count	Average	Measure Model Logit	Error	Infit MnSq	Std	Outfit MnSq	Std	Nu	Item	
211	259	0.8	-2.28	0.16	1.0	0	0.9	0	1	Q4	(V_A)
7	259	0.0	2.97	0.38	1.0	0	0.9	0	2	Q5	(Prep)
15	259	0.1	2.16	0.27	1.0	0	1.0	0	3	Q6	(PhV)
133	259	0.5	-0.78	0.13	1.0	0	1.0	0	4	Q7	(V_T)
138	259	0.5	-0.86	0.13	1.0	0	1.0	0	5	Q8	(V_T)
68	259	0.3	0.36	0.14	1.0	0	1.0	0	6	Q9	(Prep)
70	259	0.3	0.32	0.14	1.0	0	1.0	0	7	Q10	(N_A)
85	259	0.3	0.03	0.14	1.0	0	1.0	0	8	Q11	(N_N)
75	259	0.3	0.22	0.14	1.0	0	1.0	0	9	Q12	(V_F)
185	259	0.7	-1.69	0.14	0.9	0	0.9	0	10	Q13	(Adv)
61	259	0.2	0.51	0.15	1.0	0	1.0	0	11	Q14	(Prep)
226	259	0.9	-2.73	0.19	1.0	0	0.9	0	12	Q15	(V_A)
20	259	0.1	1.85	0.23	1.0	0	1.0	0	13	Q16	(Prep)
103	259	0.4	-0.28	0.13	1.0	0	1.0	0	14	Q17	(N_N)
140	259	0.5	-0.89	0.13	1.0	-1	1.0	-1	15	Q18	(V_F)
176	259	0.7	-1.51	0.14	0.9	-1	0.9	-1	16	Q19	(V_T)
127	259	0.5	-0.68	0.13	1.0	0	1.0	0	17	Q20	(V_A)
3	259	0.0	3.83	0.58	1.0	0	1.3	0	18	Q21	(Prep)
22	259	0.1	1.75	0.22	1.0	0	1.1	0	19	Q24	(PhV)
60	259	0.2	0.54	0.15	1.0	0	1.1	0	20	Q25	(Adv)
197	259	0.8	-1.94	0.15	0.9	0	0.9	-1	21	Q26	(Art)
191	259	0.7	-1.81	0.14	1.0	0	1.0	0	22	Q27	(Prep)
186	259	0.7	-1.71	0.14	1.0	0	1.0	0	23	Q28	(N_N)
42	259	0.2	0.99	0.17	1.0	0	1.0	0	24	Q29	(Prep)
16	259	0.1	2.09	0.26	1.0	0	0.9	0	25	Q30	(Art)
109	259	0.4	-0.39	0.13	1.0	0	1.0	0	26	Q31	(V_T)
118	259	0.5	-0.53	0.13	1.0	0	1.0	1	27	Q32	(N_N)
68	259	0.3	0.36	0.14	0.9	0	0.9	-1	28	Q33	(V_F)
17	259	0.1	2.03	0.25	1.0	0	1.0	0	29	Q34	(Conj)
184	259	0.7	-1.67	0.14	1.0	0	1.0	0	30	Q35	(V_F)
163	259	0.6	-1.28	0.13	1.0	0	1.0	0	31	Q36	(N_N)
93	259	0.4	-0.11	0.13	1.1	1	1.1	2	32	Q37	(Art)
215	259	0.8	-2.39	0.17	1.0	0	0.9	0	33	Q38	(Spell)
10	259	0.0	2.59	0.32	1.0	0	1.0	0	34	Q39	(N_F)
23	259	0.1	1.70	0.22	1.0	0	1.1	0	35	Q40	(Perp)
40	259	0.2	1.05	0.17	1.0	0	1.1	0	36	Q41	(Pr_A)
156	259	0.6	-1.16	0.13	1.0	0	1.0	0	37	Q42	(V_F)
144	259	0.6	-0.96	0.13	1.1	2	1.1	2	38	Q43	(Art)
34	259	0.1	1.25	0.19	1.0	0	1.1	0	39	Q44	(N_N)
198	259	0.8	-1.96	0.15	1.0	0	1.0	0	40	Q45	(V_F)

84	259	0.3	0.05	0.14	1.0	0	1.0	0	41	Q47	(Pr)
240	259	0.9	-3.36	0.24	1.0	0	1.2	0	42	Q48	(Art)
15	259	0.1	2.16	0.27	1.0	0	1.0	0	43	Q49	(Prep)
151	259	0.6	-1.08	0.13	1.0	0	1.0	0	44	Q50	(N_N)
7	259	0.0	2.97	0.38	1.0	0	1.1	0	45	Q52	(Prep)
156	259	0.6	-1.16	0.13	1.0	0	1.0	0	46	Q53	(Art)
89	259	0.3	-0.57	0.13	1.0	0	1.1	1	47	Q54	(often)
249	259	1.0	-4.58	0.32	1.0	0	1.0	0	48	Q55	(that)
10	259	0.0	2.06	0.32	1.0	0	0.8	0	49	Q56	(those)
222	259	0.9	-3.13	0.18	1.0	0	1.1	0	50	Q57	(mainly)
0	259		Maximum						51	Q58	(whether)
98	259	0.4	-0.73	0.13	1.0	0	1.0	0	52	Q59	(without)
75	259	0.3	-0.31	0.14	1.0	0	1.0	0	53	Q60	(them)
206	259	0.8	-2.68	0.16	1.0	0	1.1	1	54	Q61	(their)
133	259	0.5	-1.31	0.13	1.0	0	1.0	0	55	Q62	(in)
45	259	0.2	0.38	0.17	1.0	0	0.9	0	56	Q63	(and)
35	259	0.1	0.68	0.18	1.0	0	0.9	0	57	Q64	(as)
54	259	0.2	0.14	0.16	1.0	0	0.9	0	58	Q65	(rather)
61	259	0.2	-0.02	0.15	1.0	0	1.0	0	59	Q66	(it)
16	259	0.1	1.56	0.26	1.0	0	1.0	0	60	Q67	(not)
1	259	0.0	4.41	1.00	1.0	0	0.7	0	61	Q68	(instead)
32	259	0.1	0.79	0.19	1.0	0	0.9	0	62	Q69	(when)
116	259	0.4	-1.03	0.13	1.1	2	1.1	1	63	Q70	(then)
4	259	0.0	3.01	0.50	1.0	0	1.0	0	64	Q71	(its)
96	259	0.4	-0.70	0.13	1.0	0	1.0	0	65	Q72	(this)
35	259	0.1	0.68	0.18	1.0	0	1.0	0	66	Q73	(where)
101	259	0.4	-0.78	0.13	1.0	0	1.0	0	67	Q74	(by)
1	259	0.0	4.41	1.00	1.0	0	0.5	0	68	Q75	(whose)
89	259	0.3	-0.57	0.13	1.0	0	1.0	0	69	Q76	(sometimes)
9	259	0.0	2.17	0.34	1.0	0	1.5	1	70	Q77	(they)
23	259	0.1	1.16	0.22	1.0	0	1.2	0	71	Q78	(then)
108	259	0.4	-0.90	0.13	0.9	-1	0.9	-2	72	Q79	(who)
21	259	0.1	1.27	0.23	1.0	0	1.1	0	73	Q80	(credits)
15	259	0.1	1.63	0.27	1.0	0	0.8	0	74	Q81	(that)
45	259	0.2	0.38	0.17	1.0	0	1.0	0	75	Q82	(by)
38	259	0.1	0.58	0.18	1.0	0	1.1	0	76	Q83	(that)
69	259	0.3	-0.19	0.14	1.0	0	1.1	0	77	Q84	(those)
59	259	0.2	0.03	0.15	0.9	0	0.9	-1	78	Q85	(those)
41	259	0.2	0.49	0.17	1.0	0	1.1	0	79	Q86	(but)
103	259	0.4	-0.82	0.1	1.0	0	1.0	0	80	Q87	(and)
214	259	0.8	-2.89	0.17	1.0	0	1.0	0	81	Q88	(than)
72	259	0.3	-0.25	0.14	1.0	0	1.0	0	82	Q89	(better)
180	259	0.7	-2.12	0.14	1.0	0	1.0	0	83	Q90	(like)
140	259	0.5	-1.43	0.13	1.0	1	1.0	1	84	Q91	(now)
192	259	0.7	-2.36	0.15	1.0	0	1.0	0	85	Q92	(and)
49	259	0.2	0.27	0.16	1.0	0	1.0	0	86	Q93	(though)
31	259	0.1	0.82	0.19	1.0	0	1.0	0	87	Q94	(they)
123	259	0.5	-1.15	0.13	0.9	-2	0.9	-2	88	Q95	(but)
23	259	0.1	1.16	0.22	1.0	0	1.0	0	89	Q96	(so)
25	259	0.1	1.07	0.21	1.0	0	0.9	0	90	Q97	(also)
30	259	0.1	0.86	0.20	1.0	0	1.0	0	91	Q98	(one)
187	259	0.7	-2.26	0.14	0.9	-1	0.9	-1	92	Q99	(as)
0	259		Maximum						93	Q100	(provided)
34	259	0.1	0.72	0.19	1.0	0	0.9	0	94	Q101	(those)
32	259	0.1	0.79	0.19	1.0	0	0.9	0	95	Q102	(only)
35	259	0.1	0.68	0.18	1.0	0	1.0	0	96	Q103	(it)

			Measure	Model	Infit		Outfit			
Score	Count	Average	Logit	Error	MnSq	Std	MnSq	Std	Nu	Item

88.0	259.0	0.3	-0.00	0.20	1.0	0.1	1.0	0.1	Mean of Count:	
70.2	0.0	0.3	1.71	0.14	0.0	0.6	0.1	0.8	S.D.	96

RMSE 0.25 Adj S.D. 1.70 Separation 6.86 Reliability 0.98
Fixed (all same) chi-square: 5237.64 d.f.: 93 significance: .00

Table 1: Item Measurement Report (ordered by N).

The range of item difficulties covered extend from logit -4.58 (Item 48) to 4.41 (Items 61 & 68). Most of the items are accepted by the model with the exception of Item 18 (Infit: 1.0, Outfit 1.3), Item 42 (Infit: 1.0, Outfit: 1.2), Item 70 (Infit:1.0, Outfit: 1.5), Items 51 and 93 have been answered correctly by none. The test has thus ninety-one items accepted by the model with a fairly wide range of difficulty levels.
FACETS also reports test of the overall calibration of a facet. These are found at the bottom of the table. ('RMSE' is the root mean square standard error; 'Adj S.D.' is the standard deviation of the estimates after removing measurement error; 'Separation' is a measure of the relative spread of the estimates; 'Reliability' is the Rasch equivalent to the KR-20 or Cronbach Alpha statistics. 'Fixed chi-square' is the goodness-of-fit test for the elements' sharing the same measure after allowing for measurment error.) In the case of the item calibration, the differences (separation) among the items are found to be realiably distinct (reliability: 0.98) and the measurement variable established is consistent.
C. The Sub-tests
Table 2 reports the calibration of the two sections of the test. Part 2 (cohesion features - logit 0.14) is more difficult than Part 1 (grammar features - logit -0.14). The fit statistics are all within the accepability level. The two sections are also reliably distinct (reliability: 0.97) and the measuring variable consistent.

			Measure	Model	Infit		Outfit		
Score	Count	Average	Logit	Error	MnSq	Std	MnSq	Std	Part
4782	11914	0.4	-0.14	0.02	1.0	0	1.0	0	Part 1
3666	12950	0.3	0.14	0.02	1.0	0	0.9	0	Part 2
4224.0	12432.0	0.3	0.00	0.02	1.0	0.1	1.0	-0.2	Mean
558.0	518.0	0.1	0.14	0.00	0.0	0.3	0.0	0.5	S.D. 2

RMSE 0.02 Adj S.D. 0.14 Separation 6.05 Reliability 0.97
Fixed (all same) chi-square: 75.11 d.f.: 1 significance: .00

Table 2: Sub-test Measurement Report
D. The Student Facet
Owing to the large number of candidates it is not practicable to include a detailed person measurement report in the paper. The overall range of candidate ability is between logit -1.15 to 1.42. Table 3 reports the calibration report of the two candidates sub-groups: 'Student' and 'Reference'. The 'Reference' group (logit: -0.93) is calibrated higher than the 'Student' group (lofit: -1.16) with a reliability of separation at 0.92 and a significant overall measurement fit.

Score	Count	Average	Calib Logit	Model Error	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Group
7021	21216	0.3	-1.16	0.02	1.0	0	1.0	0	Student
1427	3648	0.4	-0.93	0.04	1.0	0	1.0	0	Reference
4224.0	12432.0	0.4	-1.04	0.03	1.0	0.1	1.0	-0.4	Mean (Count)
2797.0	8784.0	0.0	0.11	0.01	0.0	0.2	0.0	0.2	S.D. 2

RMSE 0.03 Adj S.D. 0.11 Separation 3.31 Reliability 0.92
Fixed (all same) chi-square: 23.94 d.f.: 1 significance: .00

Table 3: Group Measurement Report (ordered by N).

VII. Discussion

A. The principal research question in the study: the establishment of equivalence between the 'Reference' and the 'Student' groups has been achieved through the employment of FACETS. The logit level of the 'Reference' group (-0.93) can be taken as the equivalence of the minimal required ESL level for entry into university. The concept of equivalence should be correctly understood. Equivalence here refers to the two groups of candidates on the basis of the test administered. It does not refer to the EAP test and the Use of English examinations. Thus, while the two groups of candidates have been compared regarding ESL ability, they have not been compared regarding possible equivalence in the results of the Use of English examinations.

B. The calibration of the two parts of the test is interesting in that it enables analysis of groupings of test items. The analysis reported is in fact a construct validation study as suggested by Wright & Masters (1982:93):

> "The pattern of item calibration provides a description of the reach and hierarchy of the variable. This pattern can be compared with the intentions of the item writers to see if it confirms their expectations concerning the variable they wanted to construct. To the extent that it does, it affirms the construct validity of the variable."

The finding that cohesion features require a more advanced (difficult) ESL ability to master than grammar seems to confirm applied linguistic and TESL theory, and the views of many TESL colleagues.

As an item oriented technique Rasch analysis can be used for item oriented construct validatiaon (eg. Lee 1992). As FACETS allows for facets of item sub-groups to be included in the analysis, construct validation can also be carried out on item sub-groups. In the study reported it may not be very instructive to estimate the construct validaty directly from the items. Using the groupings of the items as it has been done would make more sense in terms of both computation and applied linguistic theory.

VII. Concusion and implications

What the study has shown are possible exetnsions of the Rasch model in terms of both item calibration and person measurement through the employment of FACETS. Indeed the package allows for a maximum of nine facets to be calibrated simultaneously. Such extensions are particularly attractive to those colleagues who, while appreciating the rigour in measurement offered by the Rasch model, would be apprehensive of the danger of being strait-jacketed in their applied linguistic research. What has been demonstrated in the

paper is that FACETS is able to maintain the rigour of the Rasch model and to provide the applied linguist with intersting research design possibilities. By so doing, FACETS has outgrown the Rasch model from being a strictly measurement model to a general research tool and enables language testers to "devote their creative powers to designing tests which involve deeper and more relevant evidence of competence..." (Linacre 1989b:10).

REFERENCES

Bachman, L. 1990. Fundamental Considerations in Language Testing. OUP.

Choi, I.C. & Bachman, L. 1992. An investigation into the adequacy of three IRT models for data from two ESL reading tests. Language Testing: 9,1:30-50.

Griffin, P.E. 1985. The Use of Latent Trait Models in the Calibration of Tests of Spoken Language in Large-Scale Selection-placement Programs. Lee et al. eds. New Directions in Language Testing: 149-162.

Henning, G. 1984 Advantages of latent trait measurement in language testing, Language Testing 1,2: 123-33.

Henning, G. 1992. Dimensionality and construct validity of language tests. Language Testing 9,1:1-11.

Henning, G., Hudson, T. & Turner, J. 1985 Item response theory and the assumption of unidimentionality for language tests, Language Testing 2,2: 141-154.

Ingram, D.E. & Wylie, E. 1979 Australian Second Language Proficiency Ratings. Adult migrant education program teachers manual. Canberra: Department of Immigration and Ethnic Affairs. Reprinted 1984 Canverra: Australian Government Publishing Service.

Jensen, A.R. 1978. g: Outmoded Theory or Unconquered Frontier?, Creative Science & Technology 2,3: 16-29.

Lee, Y.P. 1992 Language Test Validation via Rasch Analysis. Occasional Papers in Applied Language Studies: 57-66.

Lee, Y.P., Wylie, E., McKay, P. & Ingram, D.E. forthcoming. Process-Oriented Language Assessment.

Linacre, M. 1989a. Many-faceted Rasch measurement. Chicago: MESA Press.

Linacre, M. 1989b. Constructing measurement from judge-awarded ratings wiht a Many-Facet Rasch model. U. of Chicago.

Linacre, M. & Wright, B.D.1990. FACETS. MESA Press: Chicago.

MacKay, P., Hudson, C. & Sapuppo, M. 1992. The NLLIA ESL Bandscales. NLLIA ESL Development: Language and Literacy in Schools Project Report Vol.1. Canberra: national Languages and Literacy Institute of Australia.

Pollitt, A. & Hutchinson, C. 1987 Calibrating graded assessments: Rasch partial credit analysis of performance in writing, Language Testing 4,1: 72-92.

Reckase, M.D. 1979. Unifactor Latent Trait Models Applied to Multifactor Tests. Journal of Educational Statistics 4: 207-30.

Woods, A. & Baker, R. 1985 Item response theory, Language Testing 2,2: 117-140.

Wright, B.D. & Masters, G.N. 1982 <MI>Ratinqg Scale Analysis<D>, Chicago: Mesa Press.

Appendix: the EAP Test

What (1)peoples buy today, they throw away tomorrow. But (2)find somewhere to put the rubbish is (3)become harder and more expensive. America's Environmental Protection Agency (4)estimate that 80% of the country's landfills will shut (5)in 2010. Japan looks (6)^ running out of usable space by 2005. Holland has more or less (7)runed out already. Other options are no easier. Most industrial countries (8)agree two years ago to discourage shipment of hazardous waste (9)for the third world. No wonder that (10)rich-countries governments consider waste (11)disposals as their most (12)pressed environmental problem.
The problem is (13)large manmade. Rarely is there an absolute shortage (14)in space to put more rubbish dumps. But nobody (15)want a dump, or an incinerator, (16)in next door. So the piles of waste grow, while the places to pile them diminish. This affects (17)company in two ways. First, (18)get rid of hazardous waste is (19)become more expensive. This is partly because landfill (20)cost have soared; and also because companies now face lengthy paper-chase, filling (21)^ forms that record every stag of their (22)waste progress, from factory gate to (23)the dump. As a result, more and more companies disposel (24)^ their own hazardous waste; or they (25)(expensive) change the way they work so as to reduce, (26)^ amount they create. Secondly, the difficulty, (27)^ getting rid of ordinary household rubbish is driving some (28)government to impose new obligations (29)to companies, making them take back their products when (30)^ customer wants to be rid of them. That in turn is (31)change the way companies design (32)product like computers and cars.

Government 33 ^ caught between voters, who do not want more dumps 34 or incinerators, and consumers, who want to go on 35 buy things that will one day be rubbish. Confronted by the incompatible wishes of each 36 citizens, governments oten expect companies to provide 37 ^ answers. Sometimes this is sensible, but not 38 alway. One grand priece of foolishness: most of 39 America federal environmental spending goes on the pursuit 40 for companies that once dumped hazardous waste (usually legally), to make 41 it payfor 42 clean up old sites. So far, it is mainly lawyers who have cleaned up. When the law has not been broken, 43 ^ cost of clearing old 44 dump ought to be carried by the taxpayer. As for new waste, the cost of getting rid of it should 45 rests on 46 ^ companies that create it.
47 Other piece of foolishness: the unquestioning assumption that recycling is 48 ^ best way to reduce the mounds of municipal rubbish. This belief starts 49 in a self-evident truth -- that if 50 bottle and tins can lead a second life, there will be 51 fewer waste, But the argument is then taken 52 ^ irrational lengths, with governments setting targets for 53 ^ amount of an industry's product that has to be reused.

Recycling is sometimes an efficient solution; 54 ____ it is not. The materials 55____ are easiest to recycle (such as aluminium cans) are rarely newspapers and companies, rarely 56____ which bulk largest in landfills (57 ____ newspaper and directories). Recycling schemes, 58____ run by towns or by companies, rarely work 59___ subsidy; the only way to make 60___ economically self-supporting is to create a steady demand for 61___ final product. Difficult in theory, impossible 62____ practice: markets for raw materials are notoriously unsteady, 63____ that is as true for recycled pulp and plastic 64____

for cocoa and chemicals.

65____ than subsidising one solution, or bullying companies into adopting 66____, governments need to tackle the root causes of the municipal-rubbish mountain. Most goods in short supply become increasingly expensive, warning people to change their ways. That is 67____ so with rubbish disposal. People pay nothing to throw away an extra piece of trash.68____, the old newspapers and bottles in the rubbish bin magically vanish 69____ the dustmen cart them away. The first goal for policy should be to make polluters carry the true financial and environmental costs of waste disposal, and 70____ leave them to decide the most efficient response.

One good way to induce companies to cut waste is to set in industry a national target for 71____ contribution to the waste stream, and leave companies to decide how best to meet it. 72____ has been the approach in Holland, 73____ industries have accepted a goal of cutting packaging by 10% 74____ the end of the century; and in France, 75____ environment minister has asked industry to come up with ideas to cut waste sharply by the end of the century. Best of all would be to allocate companies quotas (76____ called "credits") for the amount of waste 77____ contributed to the nation's bins; and 78____ encourage those 79____ reduced most cheaply their share of rubbish to sell off spare 80____ to those who found it more costly to cut back. A variation on 81____ idea -- suggested by Project 88, an American public-policy study -- would encourage newspapers to use more recycled fibre 82____ setting a national target, and then allowing papers 83____ beat it to sell their spare "share" to others that failed to meet it.

Lots of countries try to coax people to return bottles by insisting on a refundable deposit. 84____ schemes strike many people as fair: they tax only 85____ who chuck the bottle away. 86 ____ the size of the deposit, 87 ___ the costs of administering the scheme, are generally far greater 88 ___
the environmental damage caused by discarded bottles. It would be 89___ to save such tactics for those really hazardous items which people sometimes dump in dustbins and ditches: 90___ the lead-acid battery, 91___ the main source of lead in America's environment. Some American states, including Maine 92___ Phode Island, find deposits on car batteries encourage people to bring them back -- 93 if refundable deposits are set too high, 94___ encourage naughty people to steal batteries.
In most countries the supply of rubbish is growing 95___ the supply of rubbish dumps are shrinking. 96___ it is not enough to reduce the supply of waste; governments 97___ need to increase the supply of sites. 98___ way may be to encourage local people to see these sites 99___ a source of income. 100___ tough safety rules are set and policed, cities and states could look for ways to reward directly 101___ who agreed to live near an incinerator or a waste tip. Getting rid of other people's rubbish has always been a perfectly respectable way to earn a living. 102___ when modern societies start putting a value on 103___ will they realise just how much it is worth.